

AGI: A Linguistic Blueprint

Toward a Structurally Grounded Model of Meaning Using Arabic Morphology

Author: Saif Allah Mathouthi

Affiliation: Independent Researcher

Date: 26 January 2026

Abstract

Contemporary artificial intelligence systems, particularly large language models (LLMs), achieve impressive fluency through large-scale statistical learning but remain fundamentally limited in their capacity for systematic semantic understanding, internal verification, and logical grounding. This paper explores an alternative research direction: the construction of a structurally grounded semantic inference engine inspired by the derivational morphology of the Arabic language.

The central hypothesis is not that Arabic is intrinsically superior, nor that linguistic structure alone is sufficient to yield Artificial General Intelligence (AGI), but that Arabic morphology constitutes a rare, well-developed natural system in which form–meaning relationships are explicitly regulated. Such regulation makes it a suitable empirical laboratory for investigating whether meaning can be computationally modeled as a composition of abstract semantic primitives and formally defined operators.

We propose a layered architecture consisting of (1) a root-based semantic database, (2) a formally encoded set of morphological templates acting as semantic operators, and (3) a deterministic structural parser capable of decomposing surface forms into their generative components. We illustrate how this system can infer meanings of previously unseen words without reliance on statistical prediction. We then analyze the limitations, unresolved challenges, and epistemic risks of the approach, and explicitly distinguish what this model can plausibly achieve from what remains unknown.

1. Introduction

1.1 Background and Motivation

The dominant paradigm in modern artificial intelligence relies on statistical learning from large corpora. Large language models operate by estimating the probability of token sequences conditioned on prior context. While this approach has produced systems capable of generating fluent and contextually appropriate text, it exhibits persistent weaknesses:

The absence of explicit internal representations of meaning.

The inability to verify semantic consistency independently of context.

The phenomenon commonly referred to as “hallucination,” where syntactically coherent outputs lack grounding or internal justification.

These issues do not necessarily arise from insufficient data or computational power. Rather, they reflect a deeper architectural limitation: meaning is implicitly approximated rather than explicitly represented.

1.2 Research Question

This paper investigates the following question:

Can a formally structured natural language system be used as a reference model for building a computational architecture in which meaning is generated, constrained, and validated through explicit internal structure rather than statistical correlation?

Arabic morphology is selected as a case study because it provides a naturally occurring, highly regular system of derivation in which semantic transformation is systematically encoded in morphological form.

1.3 Scope and Non-Claims

It is essential to state clearly what this paper does not claim:

It does not claim to solve AGI.

It does not claim that Arabic is uniquely capable of supporting intelligence.

It does not claim zero hallucination as an absolute guarantee.

It does not claim superiority over neural methods in all tasks.

Instead, this paper presents a research blueprint: a precise, testable, and falsifiable architectural proposal.

2. Linguistic Foundations

2.1 Root–Pattern Morphology

Arabic words are typically generated through the interaction of:

A root (usually trilateral or quadrilateral), encoding an abstract semantic field.

A morphological pattern (وزن), encoding grammatical and semantic transformations.

For example:

Root: خ-ر-ج

Abstract semantic field: emergence / outward movement.

Pattern: استفاعل (istaf'ala)

Semantic operator: seeking, requesting, or initiating an action.

The resulting word “استخرج (to extract)” is not an arbitrary string but a structured semantic composition.

2.2 Structural Properties Relevant to Computation

Several properties make this system computationally interesting:

Generativity: A finite set of roots and patterns produces a large lexical space.

Compositionality: Meaning emerges from the combination of stable components.

Constraint: Not all combinations are valid; illegal forms can be detected structurally.

Predictability: Novel but well-formed words are often intelligible to native speakers.

These properties suggest that Arabic morphology may be formalizable as a semantic algebra.

3. Conceptual Architecture

3.1 Overview

We propose a three-layer architecture:

Semantic Root Database

Morphological Template Layer

Structural Parser and Inference Engine

This system is intentionally not a black box.

4. Layer I: The Semantic Root Database

4.1 Purpose

The root database serves as the system's foundational semantic substrate.

4.2 Root Inventory

While the theoretical combinatorial space of Arabic roots exceeds 600,000 possibilities, empirical evidence shows that only a small subset is attested. Therefore:

Only attested roots extracted from authoritative lexica and corpora are included.

Each root is assigned a semantic attribute set, not a textual gloss.

4.3 Semantic Representation

Instead of natural-language definitions, each root is represented as a structured attribute vector, for example:

Copy code

Root: خ-ر-ج

Attributes:

- Motion(outward)
- Boundary-crossing
- State-change

This representation is intentionally minimal and revisable.

4.4 Epistemic Humility

Assigning a “core meaning” to a root is inherently uncertain.

All entries therefore include:

Confidence scores

Source references

Revision metadata

5. Layer II: Morphological Templates as Semantic Operators

5.1 Templates as Functions

Each morphological pattern is modeled as a deterministic function:

Copy code

Template(root_attributes) → augmented_attributes

Example:

Copy code

Pattern: استنفعِل

Operator effects:

- Add(Intentionality)
- Add(Request/Seeking)

5.2 Implementation Considerations

Templates are encoded as pure functions (e.g., Python or FST-based).

No template directly “creates meaning”; it only transforms existing attributes.

5.3 Limitations

Not all semantic effects of templates are uniform across all roots.

Exceptions and idiomatic shifts are expected and explicitly acknowledged.

6. Layer III: Structural Parser and Inference Engine

6.1 Morphological Decomposition

Given a surface form, the parser attempts to recover:

Root

Template

Affixes

Morphosyntactic features

Example:

Input: مستخرج

Decomposition:

Prefixes: مـسـتـ

Pattern: مستفعل

Root: خـرـجـ

6.2 Logical Fusion

The system performs structured composition:

Copy code

Root semantics

+ Template semantics

+ Grammatical role

→ Inferred concept

Result:

“An agent who performs the act of extracting.”

This inference is explainable and traceable.

7. Worked Example: Novel Word Interpretation

Suppose the system encounters a previously unseen word:

مستزرع

Steps:

Decomposition:

Root:) ع-ز-ع (growth, planting)

Pattern: مستفعل

Retrieval:

Root attributes: biological growth, insertion into substrate

Template attributes: intentional seeking

Fusion:

Resulting inference: “an entity that performs or seeks cultivation”

The system reaches this understanding without prior exposure.

8. Evaluation Criteria

The system must be evaluated empirically, not rhetorically.

Metrics include:

Morphological parsing accuracy

Semantic inference precision

Explainability (trace completeness)

Failure transparency

Energy cost per inference

No claim of superiority is meaningful without such benchmarks.

9. Limitations and Open Problems

9.1 Semantic Underdetermination

Roots are often polysemous. No representation fully resolves this.

9.2 Metaphor and Pragmatics

Many meanings arise from discourse, not morphology.

9.3 Coverage

A symbolic system fails where knowledge is absent.

9.4 AGI Gap

General intelligence involves perception, embodiment, learning, and goals.

This architecture addresses semantic grounding only.

10. Conclusion

This paper proposes a structurally grounded approach to semantic modeling inspired by Arabic morphology. It does not offer a shortcut to AGI, nor a replacement for statistical learning. Instead, it contributes a concrete, testable hypothesis:

Explicit structural constraints on meaning may reduce semantic arbitrariness and improve interpretability in artificial systems.

Arabic morphology serves here not as a cultural artifact, but as an empirical model of regulated meaning generation. Whether such a model can scale, generalize, or integrate into broader intelligent systems remains an open question—one that must be answered experimentally, not philosophically.